A Machine Learning Model for Prediction of Heart Disease

अ मशीन लर्निंग मॉडल फॉर प्रेडिक्शन ऑफ़ हार्ट डिसीज

A

SYNOPSIS

SUBMITTED TO

CHAUDHARY DEVI LAL UNIVERSITY, SIRSA

For Registration

Of

DOCTOR OF PHILOSOPHY

In

Computer Science and Applications

Supervisor:

Prof. Dilbag Singh

Investigator:

Jasjit Singh Samagh

Department of Computer Science & Applications
Chaudhary Devi Lal University, Sirsa-125055 (India)

2018-19

INDEX

1. Introduction

1.1 Machine Learning: An outline

1.2 Machine Learning Approaches

1.3 Machine Learning Classification Techniques

2. Literature Review

2.1 Research Gap Analysis

3. Research Methodology

3.1 Problem Statement

3.2 Research Objectives

3.3 Significance of the Study

3.4 Chapterization

4. References

1. INTRODUCTION

Cardiovascular disease in general relate to situations that involve contracted or blocked blood vessels that can lead to heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect hearts muscle, valve or rhythm also are considered form of heart disease. Heart diseases are also called as silent killers because symptoms are very difficult to detect. Heart diseases are increasing day by day. As per World Health Organization calculations over 17 million people die every year due t heart disease like heart attacks and strokes therefore it is necessary to have a record of crucial evidence and health problems that lead to Heart disease. If the heart disease found in born babies, then it is called congenital Heart Disease and if it is at the later ages, then it is called acquired Heart Disease. For diagnosis of heart disease number of tests are performed namely blood pressure, ECG, auscultation, cholesterol and blood sugar. These tests procedures are long and consume a lot of time and administering of medicines gets delayed which sometimes may have adverse effect on the patient. Machine learning with its tools can help doctors, pathologists to reduce the timing of such tests. It further helps them to be more accurate because amount of data is increased. With machine learning it is very easy to get knowledge from a large data which is not possible for man to analyse [5].

1.1 MACHINE LEARNING: AN OUTLINE

In 1959, the term Machine learning was given by Arthur Samuel which defines Machine learning as a discipline which gives computers the power to discover with not being programmed explicitly [7]

Machine learning is the process to generate the algorithms based on the previous inputs and past experiences. [7]

To understand the use of machine learning more accurately we will consider some of the examples where the concept machine learning is applied. Such examples are self-driving car, Google listening search, dashboards on Amazon, showcases on Netflix and many more are the examples of machine learning. [7]

Let us understand machine learning with one more example now just imagine that you are using a news application, but the application is not able to provide you with latest notifications and

updates on time. And now this will make you think that the application is of no use and it is not making any aid to me. So, in this case the concept of machine learning is used through which we can achieve the best from the application or software. Machine learning uses the concept that enables the automatic notification feature provider by the application. This is a very common real-life example of why we need the concept like machine learning. [7]

1.2 MACHINE LEARNING APPROACHES

For solving a problem basically four types of Machine Learning approaches being used they are supervised, unsupervised, semi supervised and reinforcement learning. The nature of data labelling is vital for Machine Learning. The first three approaches basically use data labelling strategy for solving a problem, and the fourth and the last category is the Reinforcement learning model that basically interacts with environment.[4]

Supervised Learning

Supervised learning is the machine learning approach that consists of techniques that learn from externally supplied examples that results in a general theory which makes prediction about future instances. An outcome or output-variable guides the whole learning process. A learning process assesses an unknown sample from known sample, where sample cold be both input or output and where output is labelled as classification or regression. The process employs a known dataset called training dataset that enables predictions. A training dataset consists of input data and response values. Learning process seeks to build a model from the dataset that can make predictions of the response values for a new dataset. A test dataset is often used to confirm the performance of a Machine Learning model. Larger training datasets often produces the models with greater predictive power that can better simplify new datasets. Examples of supervised learning algorithms are Decision tree, KNN, SVM, ANN, and Random Forest etc.[4]

Unsupervised Learning

Unsupervised learning is a type of machine learning approach which is used to draw conclusions from datasets which consist of input data with no labelled responses. In this approach there are basically two learning tasks namely Association and clustering. To find the correlations in the objects of a database, Association learning was proposed by Rakesh Agarwal. The most familiar

algorithm used in association rule is Apriori and clustering is used to group same type of datasets. Some of the most common algorithms used in unsupervised learning approach are k-means clustering and association rule learning algorithm.[5]

Semi-supervised Learning

Semi-supervised learning is a machine learning approach which is basically the combination of supervised and unsupervised learning, where portion of data is partially labelled. This learning is basically used in classification of webpage, genetic sequencing and speech recognition. Semi-supervised learning is broadly classified into two learning tasks namely classification and clustering.[5]

Reinforcement Learning

Reinforcement Learning is one of the Machine Learning method which deals with how software agents will be used automatically to determine the perfect activity of a specific context so as to maximize its performance. Reinforcement signal sends the reward feedback for the agent to learn its behaviour. It consists of two learning tasks namely classification and control. Some applications of the reinforcement learning are computer played board games, robotic hands, and self-driving cars. Most commonly used algorithms in reinforcement learning are Q-learning, Temporal difference, Deep Adversarial Networks.[7]

1.3 ML CLASSIFICATION TECHNIQES

The Machine Learning classification techniques are as follows

Regression

Regression is one of the supervised learning model which aims to deliver the prediction of an output variable in accordance with its input and which are later identified. Most identified algorithms used are linear regression,logistic regression and stepwise regression.Many complex regression algorithms are also developed like ordinary least squares regression, multivariate adaptive regression etc [11].

Clustering

Clustering is basically used in unsupervised learning model in form of clusters to find natural groupings of data.Commonly used clustering techniques are the k-means technique, the hierarchical technique, and the expectation maximization technique [11].

Bayesian Models

Bayesian models also uses supervised learning and basically used for solving either classification or regression problems.Most commonly used algorithms are naive bayes, gaussian naive bayes, multinomial naive bayes, bayesian network, mixture of gaussians, and bayesian belief network [11].

K- Nearest Neighbor

In K-nearest neighbor (KNN) technique, nearest neighbor is measured with respect to value of k, that define how many nearest neighbors need to be examined to describe class of a sample data point. Nearest neighbor technique is divided into two categories i.e., structure based KNN and structure less KNN. The structure based technique basically uses the basic structure of the data where the structure has less mechanism and also associated with training data samples. In structure less technique entire data is categorized into sample data point and training data, distance is calculated between sample points and all training points and the point with smallest distance is known as nearest neighbor [11].

Support Vector Machines

Support vector machines (SVMs) were introduced first to be used in statistical learning theory. SVM is basically a binary classifier that create a linear separating hyper plane to sort data position.SVMs are basically used in classification, regression, and clustering.In case of global optimization SVMs deal with more complex problems which appear in high dimensional spaces which makes them attractive in various applications.Commonly used SVM algorithms are the support vector regression, least squares support vector machine, and successive projection algorithm-support vector machine [11].

Ensemble Learning

Ensemble learning models are basically used to improve the predictive performance of statistical learning technique by building a linear combination of simpler base learner. Ensemble learning models uses decision trees as the base learner, in case of random forest a large number of boosting and bagging implementations have been proposed, Ensemble Learning is basically used in boosting technique and bootstrap aggregating or bagging algorithm [11].

2. REVIEW OF LITERATURE

Youness Khourdifi et.al. (2019) compared the algorithms with different performance measures with the help of machine learning. For prediction algorithms worked differently.In this study the best results are shown by K-Nearest Neighbour, and Random Forest and Artificial Neural Network. Then he merged the algorithms and try to check the performance so as to check if it will be more efficient or not and later result is applied to heart disease data set and his proposed models resulted in better accuracy of 99.65%

Mohamed Bahaj et.al (2019) evaluated the result of proposed model and compared with the algorithms having different performance measures which resulted into better performance and effectiveness as compared to K-Nearest Neighbour, Random Forest, Naïve Bayes , Support Vector Machine SVM and Artificial Neural Network .

Komal Saini1et.al. (2019) reviewed an IoT framework for the prediction of the heart disease.Wearable sensing devices which are used to collect the data from the body of the patients and then storage and processing of data using communication standards. Among the communication standards which transfer the data collected to the storage devices, Bluetooth (BLE) has been well suited for the health care application and for the storage of the huge amount of data generated by the wearable sensor cloud storage and big data management is considered as the best. Finally, for the analysis of the data to predict a model data miming tools and machine learning algorithms are used to get the accurate results but the accuracy of all the algorithms vary according to the parameters and data collected.

SenthilKumar Mohan. et.al. (2019) stated that identifying the processing of raw health care data of heart information will help in the long term saving of human lives and early detection of

abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease.

Shadman Nashif et.al. (2018) developed system for transmission of the recorded data to a central server which are updated every 10 seconds with the help of which the doctors can see the patient's real-time sensor data by using the application and start live video streaming if instant medication is required. Other important feature of the proposed system was that if any real-time parameter of the patient exceeds the threshold then the prescribed doctor will be notified at once through GSM technology

Kusuma et.al. (2018) made a systematic effort to identify and review machine learning, data mining and deep learning approaches applied on HD research. EHR's in health care producing huge amount of data with the development of technology also give rise to in-depth exploration towards accurate disease prediction, diagnosis, and treatment. The potential benefits of applying Machine learning and deep learning methods with the suitable algorithm will reduce the disease rates and death failures of the global population. In this regard, in future state of the art performances from ML and DL applications not only in disease prediction and diagnosis but also in the other field of bio informatics.

Pooja et.al. (2018) provided different ways to implement the concept of machine learning and to get the best results from out there. Basically,machine learning is the concept to analyse the huge amount of data. In past centuries, the maximum amount of time was let's say wasted by our programmers in order to analyse the huge volume of data on their own and getting the result out of it. We all know that it took hours and hours sometimes even years to estimate the data in bulk. When at last the time of results came, then again, they get errors in achieving the desired output.

So, this was a very poor and bad approach as it consumes more time and efforts and also didn't produced the results according to the desire of the user.

Meherwar Fatima et.al. (2017) surveyed the different machine learning techniques for the diagnosis of different diseases like heart disease, diabetes disease, liver disease, dengue and hepatitis disease. He figured out several good results in his study.In previous study it is discovered that for the detection of heart disease improved accuracy of 94.60% is recorded in case of SVM. In case of diabetes disease Naive Bayes proved to diagnose correctly and accurately with accuracy of 95%.and In case of liver diagnosis 97.10% of correctness is recorded using Functional tree and In case of dengue disease detection 100% accuracy is recorded by RS theory.Author used feed forward neural network that correctly classifies hepatitis disease recorded with 98% of accuracy. In this paper writer also provided a suite of tools that are developed in community of Artificial Intelligence and are useful in the improved decision making process.

Min Chenet et.al. (2017) proposed a new disease risk prediction algorithm using structured and unstructured data which were collected from hospital. As per him there was no existing work that has focused on both data types in the area of medical big data analytic. Author compared his algorithm with already existed algorithm out of which his proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the existing algorithms.

Aized Amin Soofiet et.al. (2017) described various popular classification techniques of machine learning with their basic working mechanism, strengths and weaknesses. The solution for potential applications and issues had been highlighted. Classification methods are typically strong in modelling interactions. The classification techniques discussed can be implemented on different type of data set infield of health, business enterprise etc. To find out which technique is superior to other is very difficult because each technique is with its own merits, demerits and implementation issues. The classification techniques selection depends on user problem domain. However there is lot of work that has been done in classification domain, but it still requires formal attention of research community to overcome classification issues that have been coming to light while dealing with new classification problems as the problems in classification of Big Data.

Konstantinos G. Liakos et.al. (2017) tried to put light on the research done for the applications of machine learning in agricultural production systems. The works analyses were done on crop management which includes its applications on prediction of yield, disease detection, weed detection crop quality, and species recognition, research also includes applications on welfare of animals and also in water management and lastly in soil management. This article try to display how agriculture will be benefited from machine learning technologies.

Ahmed Fawzi Otoom et.al. (2015) proposed a system for coronary artery disease detection which can be used for analysis and followups. UCI data set is used that consists of 303 cases and 76 features out of which 13 features are used in this study.He performed two tests by using three algorithms namely Bayes Naive, Support vector machine, and Functional Trees for detection. For data mining WEKA tool is used. After testing the Holdout test that is the first test, the 88.3% accuracy is recorded by using the SVM technique. In case of the second test the cross-validation test SVM and Bayes combination is recorded with 83.8% accuracy and lastly the accuracy of 81.5 % is recorded after the use of functional tree.He therefore selected the seven best features by using the Best First selection algorithm and in case of validation, cross-validation tests are used.And later by applying the test to the selected seven best features ,Bayes Naive is recorded with 84.5% accuracy, SVM with 85.1% of accuracy and Functional Tree with 84.5% .

K.Vembandasamy et.al.(2015) used the Naive Bayes algorithm for the diagnosis of heart disease. Naive Bayes is used because it has a powerful principle of independence and and also used Bayes' theorem.All data been used in research are collected from the leading diabetes research institutes in Chennai. The data set is of 500 patients and for data mining tool used is WEKA and the result that Naive Bayes predicted is with 86.419% of accuracy.

Vikas Chaurasia et.al. (2013) suggested the use of data mining tools to detect heart disease.Author consider WEKA as the best tool as it is open source and it contains machine

learning algorithms for the purpose of mining.In his study he used Naive Bayes, J48 and bagging approaches. In this study the UCI data set is used for prediction of heart disease. Dataset includes 76 attribute but author used only 11 attributes for prediction. Naive is recorded with 82.31% accuracy. J48 is recorded with 84.35% accuracy and 85.03% of the accuracy is recorded by bagging.

G. Parthiban et.al. (2012) found the chances of heart disease from diabetic diagnosis attributes. This kind of classifiers can help in early detection of the vulnerability of a diabetic patient to heart disease. There by the patients can be forewarned to change their lifestyle. This will help diabetic patients in preventing from being affected by heart disease and thus resulting in low mortality rates as well as reduced cost on health for the state. In case of SVM it has again proved to be a best classification technique with excellent predictive performance and been investigated with the help of ROC curve for both training and testing data. Hence for the classification of the diabetic dataset SVM model can be recommended.

2.1 RESEARCH GAP ANALYSIS

The following research gaps have been identified after analyzing the relevant published literature:

- In the previous research, the performance was evaluated which was not up to the mark for the accurate prediction of Heart disease.
- The studies were only about the impacts of specific machine learning techniques and there is no information on the optimization of these techniques using optimization algorithms.
- There is no machine learning model proposed based on ensemble method and optimization algorithm for predicting the heart disease.

3. RESEARCH METHODOLOGY

A research method indicates the process in which the research will be carried out. It will specify the research methodology used in the study. In order to bridge the research gaps a systematic study is required that will propose a novel approach towards implementing machine learning in heart disease. Proposed research will be based on exploratory, experimental and applied research design. It is exploratory because it consists of discovery of new ideas and possible insight in identifying further study. Data is tested in order to test the tools, so it is experimental, and experiments conducted by designing the data mining tools. As the proposed work is for heart disease prediction, it is with the help of applied research we can easily use it in solving the problem.

3.1 PROBLEM STATEMENT

The problem is entitled as **"A Machine Learning Model for Predicting Heart Disease".** Heart disease is one of the most significant causes of death in the world today. However, if it is predicted in advance it can provide important insights to doctors for diagnosis and treatment. Thus it is necessary to have a record of health problems and crucial evidence that leads to heart diseases. Machine learning has been proved to be very effective in this predictions and impressive decisions can be drawn on the basis of dataset. To achieve this target, a machine learning model is proposed based on ensemble learning method and it is then optimized by using optimization algorithm. This model will help in improving the performance of analysis and prediction of heart diseases.

3.2 RESEARCH OBJECTIVES

1. To study the various techniques of machine learning in predicting the heart disease.
2. To identify issues and loopholes in the existing technique.
3. To propose a machine learning model based on Ensemble Learning method and optimized by optimization algorithm for predicting heart disease.
4. To implement the proposed model using Data Mining tools and dataset.
5. To compare the performance of proposed model with existing machine learning techniques.

3.3 SIGNIFICANCE OF STUDY

The research will bring a significant change in the ways the Heart Disease is being diagnosed and treated these days in this country. The optimized machine learning model using ensemble learning would help the health workers to make accurate predictions of heart disease within minimal time as it will be more efficient. The timely prediction would save the lives of those patients as they would get the accurate treatment on time.

3.4 CHAPTERIZATION

Chapter 1: Introduction

An Introduction to the Machine Learning will be elaborated in this chapter. It covers the basic understanding of machine learning, including machine learning approaches and machine learning classification techniques in detail. This chapter will also include the information about various cardiovascular diseases and the heart disease dataset to be used in predicting heart disease

Chapter 2: Review of literature

In this chapter we will review the relevant literature of the research work directly or indirectly from various sources such as reference books, journals, periodicals, research-oriented user groups and internet.

Chapter 3: Research Methodology

An insight into the research methodology will be given in this chapter. It comprises of problem statement, research objectives, scope and significance of the study, research design, analysis, chapterization and limitations of the study.

Chapter 4: Issues and Loopholes in Machine Learning Techniques

In this chapter we will be identifying and dealing with the issues and loopholes present in the existing machine learning techniques in predicting heart disease. Various machine learning techniques will be discussed with respect to current machine learning models in practice.

Chapter 5: Proposed Machine Learning Model for predicting Heart Disease

A novel machine learning model based on ensemble learning method and later optimized by optimization algorithms will be proposed for predicting the heart disease. The proposed model shall be providing solution to the issues and loopholes identified in the previous chapter.

Chapter 6: Implementation of proposed Machine Learning Model for predicting Heart Disease

This chapter will be targeted on implementation of the proposed model using data mining tools and dataset.

Chapter 7: Performance comparison of Proposed Model with existing Machine Learning Techniques

This chapter will be targeted on the comparison of the proposed hybrid machine learning model with the existing machine learning techniques in order to check the performance that which one is more efficient.

Chapter 8: Conclusion and future scope

Finally, this chapter shall be devoted to the concluding remarks and future prospects of the present work.

REFERENCES

[1] Youness Khourdifi and Mohamed Bahaji (2019). *Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization*, International Journal of Intelligent Engineering and Systems, Vol.12, No.1, 2019

[2] Youness Khourdifi and Mohamed Bahaji (2019). *The Hybrid Machine Learning Model Based on Random Forest Optimized by PSO and ACO for Predicting Heart*, ICCWCS 2019, April 24-25, Kenitra, Morocco.

[3] Komal Saini and Sandeep Sharma (2019). *Review on the Heart Disease Detection Using IoT Framework,* International Journal of Computer Sciences and Engineering Vol.7(3), Mar 2019, E-ISSN: 2347-2693.

[4] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivasatava (2019). *Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, IEEE* access volume7,2019.

[5] Shadman Nashif, Md. Rakib Raihan, Md. Rasedul Islam,and Mohammad Hasan Imam (2018).*Heart Disease Detection by Using Machine Learning Algorithms and a Real Time Cardiovascular Health Monitoring System*, World Journal of Engineering and Technology, 2018, 6, 854-873.

[6] Kusuma and Divya Udayan (2018). *Machine Learning and Deep Learning Methods in Heart Disease (HD) Research,* International Journal of Pure and Applied Mathematics Volume 119 No. 18 2018, 1483-1496.

[7] Pooja, Aakansha Sharma and Ankush Sharma (2018). *Machine Learning: A Review of Techniques of Machine Learning,* Journal of Applied Science and Computations Volume 5, Issue 7, July /2018.ISSN NO: 1076-5131.

[8] Meherwar Fatima and Maruf Pasha (2017). *Survey of Machine Learning Algorithms for Disease Diagnostic*, Journal of Intelligent Learning Systems and Applications, 2017, 9, 1-16

[9] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang (2017). *Disease Prediction by Machine Learning Over Big Data from Healthcare Communities*, IEEE Access

[10] Aized Amin Soofiand Arshad Awan (2017). *Classification Techniques in Machine Learning: Applications and Issues*, Journal of Basic & Applied Sciences, 2017, 13, 459-465

[11] Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou , Simon Pearson and Dionysis Bochtis (2018).*Machine Learning in Agriculture: A Review,*Sensors 2018, 18, 2674.

[12] Ahmed Fawzi Otoom, Emad E. Abdallah, Yousef Kilani, Ahmed Kefaye and Mohammad Ashour (2015). *Effective Diagnosis and Monitoring of Heart Disease*, International Journal of Software Engineering and Its Applications Vol. 9, No. 1 (2015), pp. 143-156.

[13] K.Vembandasamy, R.Sasipriyaand E.Deepa (2015).*Heart Diseases Detection Using Naive Bayes Algorithm*, International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 9, September 2015.

[14] Vikas Chaurasia and Saurabh Pal (2012). *Data Mining Approach to Detect Heart Diseases*, International Journal of Advanced Computer Science and Information Technology (IJACSIT)Vol. 2, No. 4, Month Year, Page: 56-66, ISSN: 2296-1739.

[15] G. Parthiban and S.K.Srivatsa (2012). *Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients*, International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868.